A Feature-Partition and Under-Sampling Based Ensemble Classifier for Web Spam Detection

Xiaoyong Lu, Musheng Chen, Jhenglong Wu, and Peichan Chan

***Abstract*—Web spam detection has become one of the top important tasks for web search engines. Web spam detection is a class imbalance problem because normal pages are far more than spam pages. However, most of traditional learning methods are not effective on imbalance classification problems. In order to tackle this problem and make full use of various features extracted from web pages' content and links, this paper presents an ensemble classifier based on under-sampling and feature-partition techniques and integrates decision tree algorithm C4.5 into it as a sub classifier to detect web spam. The experimental results show that the ensemble classifier outperforms other approaches on several evaluation metrics such as F1-Measue, AUC etc. in WEBSPAM-UK2006 dataset.**

***Index Terms*—Web spam detection, under-sampling, features partition, ensemble classifier, C4.5.**

I. Introduction

Web spam can be defined as websites or web pages which will trigger an unjustifiably favorable relevance or importance considering the page's true value [1]. The phenomenon of web spam takes place mainly due to the facts that the search engine users tend to browse only the top ranked search engine results [2], [3]. Therefore, many website owners attempt to get a higher search engine ranking by unethical ways [4]. Web spam weakens trust of users in a search engine provider, wastes an amount of computational and storage resources, deprives legitimate websites of revenue, and deteriorates the quality of search results [4]. Web spam detection has been one of the top important tasks for web search engines.

Various methods have been proposed for web spam detection. Most of these methods focus on exploring the features of distinguishing spam from normal pages. Generally, there are two major categories of features, content-based and link-based features [5], [6]. In order to promote the performance of web spam detection, machine learning techniques, including supervised, unsupervised and semi-supervised learning methods, tend to be used based on these features.

In this paper, we treat web spam detection as a binary classification problem and use supervised learning methods to classify web sites or web pages as spam or normal. There are two problems to be dealt with in web spam detection by classification methods: (1) although the number of spam pages is huge and growing, it is still the minority compared with the number of normal pages. It means that datasets for web spam detection are usually imbalance. However, most of the classification algorithms for imbalance datasets do not work well. (2) Nowadays, numerous features are extracted for web spam detection. If all of them are used to train a classifier at the same time, it will lead to over-fitting problem and decrease the performance of the classifier.

In order to resolve the problems mentioned above, we propose an ensemble classifier based on both under-sampling technique to promote the performance of the classifier for imbalance web spam datasets and features partition technique to solve the over-fitting problem. The sub classifier is C4.5 to be integrated.

Manuscript received January 5, 2015; revised July 4, 2015.

Xiaoyong Lu and Musheng Chen are with Nanchang University, China (e-mail: lxy@ncu.edu.cn, dreaminit@gmail.com).

Jhenglong Wu and Peichan Chan are with the Information Management Department, Yuan Ze University, Taiwan (e-mail: jlwu.yzu@gmail.com, iepchang@saturn.yzu.edu.tw).

II. Related Works

A. Web Spam Detection Based on Machine Learning

In general, web spam can be classified as two categories: content spam and link spam [1]. According to these two categories, two kinds of features (content-based and link-based features) can be extracted from web pages to identify spam pages [5], [6]. The features based on content are extracted from the pages' title, subject, Meta tags, anchor text, and URL etc. For example, TFIDF is a commonly used index for adversarial information retrieval. Ntoulas *et al*. extracted a series of features from web pages' content based on heuristic rules, and the experimental results on the MSN dataset showed that these features could be used to recognize web spam better [7]. The features based on links can be calculated from out-links, in-links etc. of web pages and the whole link graph of the Internet. Some of these features can be used separately such as PageRank, HITS, TrustRank, AntiTrustRank and TruncateRank etc. Castillo *et al*. took content-based and link-based features into account at the same time for web spam detection by the means of machine learning classifier [8]. They extracted more link-based features and optimized the classification algorithms to promote the performance.

In order to promote the performance of web spam detection, machine learning techniques, including supervised, unsupervised and semi-supervised learning methods, tend to be used. Scarselli *et al*. presented a cascade architecture containing a probabilistic mapping graph self-organizing map and a graph neural network to detect web spam [9]. The experiments on WEBSPAM-UK2006 showed that the results reached the state of the art when compared with some of the best results obtained by others quite different approaches. An efficient fuzzy clustering method was presented by Jegadeesh

doi: 10.18178/ijmlc.2015.5.6.551

to detect spam web pages [10]. Wang *et al*. proposed two new semi-supervised learning algorithms integrating the traditional co-training with the topological dependency based hyperlink learning to boost the performance of web spam classifiers. The experimental results showed that the algorithms are effective [11].

B. Solving Class Imbalance Problem on Web Spam Dataset

The Internet is still continuously growing. At the same time, web spam pages are also increasing dramatically. However, the number of web spam pages is still far less than the number of normal web pages. It means that datasets for web spam detection are usually imbalance. There are usually three kinds of methods to promote the performance of imbalance classification: using new evaluation metrics, changing the distribution of samples, and designing new algorithms. Studies showed that the metrics such as F-Measure, G-mean, and Weighted Accuracy etc. are more accurate to evaluate classification performance on imbalance datasets. The techniques that change the distribution of samples include over-sampling and under-sampling. Price sensitive analysis, Weighted-SVM and integrated learning methods etc. can be used as classification algorithms on imbalance datasets. Geng *et al*. proposed a novel ensemble classifier based on under-sampling technique and C4.5 decision tree classifier for web spam detection and achieved good results [12].

C. The Ensemble Classifier Based on Features Partition

Numerous features are extracted for web spam detection. If all of them are used to train a classifier at the same time, it will lead to over-fitting problem and decrease the performance of the classifier.

Peng proposed an ensemble classifier based on sub-feature space extraction to classify the micro-array datasets that contain too many characteristics [13]. Experimental results demonstrated that the classifier outperforms the classifiers generated by conventional machine learning. In this paper, features partition technique, similar to sub-feature space extraction, has been presented to integrate classifier to solve the over-fitting problem due to excessive features in web spam detection.

III. Method

In this paper, we present an ensemble classifier based on both under-sampling technique and features partition technique. In the remainder of this section, we discuss the under-sampling technique, the features partition technique, and the ensemble classifier in order.

A. Ensemble Based on Under-Sampling

Web spam datasets are imbalance datasets because non-spam pages are far more than spam pages. In this paper, we solve this problem of imbalance dataset with an ensemble method based on under-sampling technique. Under-sampling has been popularly used in class imbalance learning [14]. Given the minor example set S and the major example set *N*, under-sampling randomly samples a subset *N*' from *N*, where $|N'|<|N|$($|N|$ represents the samples number in set *N*). Since under-sampling only uses a subset of the major class examples to train the classifier, the training process is very efficient for learning algorithms that do not consider class-imbalance. However, potentially useful information contained in these ignored examples, i.e. examples in N-N', is neglected, and the neglect leads to the main deficiency of under-sampling algorithm. An ensemble strategy can be used to overcome the deficiency and keep the efficiency of under-sampling [12]-[14]. Different from [12]-[14], we set the number of *N*' approximately equal to the number of *S*, and divide all of the *N* samples into several *N*' samples subset randomly. As result, the sampling ratio $K = N'/N$ and $M=N/N'$ sub datasets are obtained.

Supposed that samples subsets N_1', N_2'…, N_m' are acquired from *N* samples, for each subset N_i', a classifier C is trained using samples N_i' and *S*. All the results generated by the sub classifiers C are combined for the final decision. Most of the classification algorithms can be adopted to be the sub classifiers. In this paper, we choose C4.5 as the sub classifier. The ensemble algorithm based on under-sampling shows as following:

Step 1: Input training set with minor class examples S and major class examples *N* (*S* and *N* correspond to spam and normal set respectively) and testing set. Supposed that the majority (Non-spam) is labeled 0 and the minority (Spam) is labeled 1.

Step 2: Initialize the spamicity of each test sample: spamicity = 0.

Step 3: Divide the majority samples *N* into *M* samples subsets N_i': {N_1'; $N2'$; … N_m'}, the number of each samples subset is approximately equal and $M = N/N'$.

Step 4: For *(*i=0; i<*M*; i++) {train the sub classifier C with the samples' subset N_i' and S and save the model as Model_i; test the testing set with Model_i and obtain the classification result CR(*x*, *C*); Spamicity=spamicity+CR(*x*, *C*);}.

Step 5: Spamicity = spamicity/*M*.

Step 6: If (spamicity >= 0.5); *X* is the minority class, Else *X* is the majority class.

The value of the *CR*(*x*, *C*) in the Step 4 can be calculated by following formula.

$$CR(x,C)=\begin{cases}1 & x \text{ is minor class}\\ 0 & x \text{ is major class}\end{cases} \quad (1)$$

where: *x* is a testing sample, *C* is a particular classifier.

B. Ensemble Based on Features Partition

The features partition refers to the division of the features set into a plurality of features subsets according to the correlation among the features. That is to say, more interrelated features are assigned to the same features subset, while less interrelated features are assigned to different features subsets. All the subsets are mutually exclusive, viz., different subsets do not have the same features. For example, features in WEBSPAM-UK2006 can be divided into 4 features subsets, shown as Fig. 1.

According to the results of features partition, the entire dataset (including training set and testing set) is longitudinally divided into several independent sub-datasets. Each of the

training subset will be used to train a sub-classifier that will be used to predict the testing samples' classification based on the features' value in the same subset. The final classification results can be gotten by integrating all the classification results obtained by different features subsets. The ensemble method based on features partition leverages the characteristics of each features subset to avoid over-fitting problem.

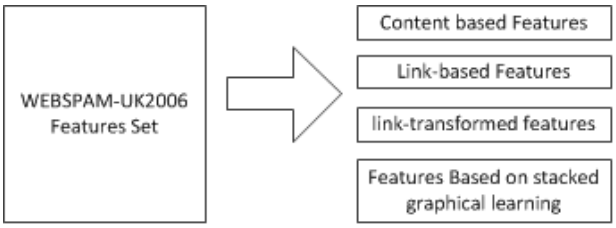

Fig. 1. Features partition in WEBSPAM-UK2006.

C. An Ensemble Classifier Based on Under-Sampling and Features Partition

By integrating under-sampling and features partition, a novel ensemble classifier can be obtained. The flow chart of the ensemble modal shows in Fig. 2. First, using the features partition technique, split all features into multiple mutually exclusive features subsets according to the correlation among features, and divides the dataset into a plurality of partial datasets according to the features partition. Second, sample each imbalance partial dataset to multiple balanced sample sets using the under-sampling method described in section 3.1. Each samples subset can be treated as a training dataset to a sub classifier. The sub classifier used in this paper is C4.5. We can predict the testing dataset and obtain the result by each sub classifier. Finally, we can get the final classification results by integrating all of the classification results.

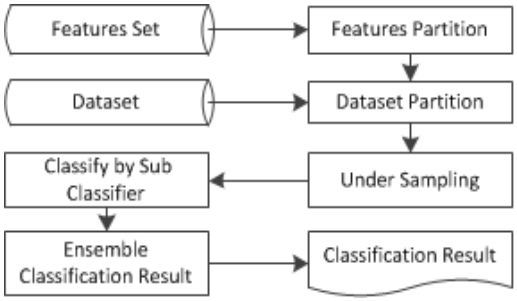

Fig. 2. Flow chart of the ensemble classifier.

IV. Experiments

A. Dataset and Evaluation Metrics

All of the experiments in our research are conducted on WEBSPAM-UK2006 [15], a publicly available collection introduced in 2007 by the Web Spam Challenge and the Adversarial Information Retrieval on the Web workshop (AIRWeb). The numbers of spam and non-spam hosts on WEBSPAM-UK2006 are displayed in Table I. We can find that the ratio between non-Spam hosts and spam hosts in the training dataset is approximately 7:1. It means that the training dataset is imbalance and it is in line with the actual situation.

There are 274 features in WEBSPAM-UK2006 dataset. All of the features can be divided into four partitions: content-based features (96), link-based features (41), link-transformed features (135), and 2 features based on stacked graphical learning.

TABLE I: WEBSPAM-UK 2006

No.	Hosts number	Spam hosts number	Non-Spam hosts number
Training set	5622	674	4948
Testing set	1851	1250	601
Total	7473	1924	5549

The experimental results are evaluated using five metrics: Accuracy, Precision, Recall, F1-Measure, and ROC AUC (the Area under the Curve of the Rece3iver Operating Characteristics) [16].

B. Experiment Results

We performed experiments on WEBSPAM-UK2006 using the under-sampling and features partition based ensemble classifier with C4.5 (C4.5+FP+US). As baselines, we also performed some experiments using C4.5, Bagging with C4.5 (C4.5+bagging), Adaboost with C4.5 (C4.5+AdaBoost), Under-sampling based ensemble classifier with C4.5 (C4.5+US), and features partition based ensemble classifier with C4.5 (C4.5+FP). All the experimental results are shown in Table II.

TABLE II: Experimental Results

Classifier	C4.5	C4.5+ bagging	C4.5+ Adaboost	C4.5+US	C4.5+FP	C4.5+ FP+US
Accuracy	0.7277	0.7763	0.7509	0.8358	0.8028	**0.8822**
Precision	0.9347	0.9514	0.9418	0.9261	0.9360	**0.8660**
Recall	0.6416	0.7048	0.6728	0.8224	0.7600	**0.9768**
F1-Measure	0.7609	0.8097	0.7849	0.8712	0.8389	**0.9180**
ROC Area	0.7742	0.8150	0.7931	0.9135	0.8902	**0.9437**

Precision, Recall and F1-Measure only denote the performance to identify the positive samples (Spam). In fact, web spam detection need to identify not only spam pages but also normal pages. AUC is a better metric to evaluate the performance in web spam detection. Comparing and analyzing the AUC results in Table II we can find that: (1) Both Bagging with C4.5 and AdaBoost with C4.5 methods improve the sub classifier's performance; (2) Either under-sampling based or features partition based ensemble classifier is better than both Bagging with C4.5 and AdaBoost with C4.5; (3) The ensemble classifier integrated under-sampling and features partition methods is the best method among the methods mentioned above.

TABLE III: Scarselli *et al.*'s Experimental Results

Classifier	FNN,PM-G,GNN(3)	GNN+GNN(1)	Autoassociator+GNN(1)	FNN,PM-G+GNN(3)+GNN(1)	C4.5+FP+US
Accuracy	0.9124	0.9070	0.9104	**0.9294**	**0.8822**
F1-Measure	0.5890	0.4400	0.4173	**0.6324**	**0.9180**
ROC Area	0.9236	0.8103	0.8070	**0.9362**	**0.9437**

Scarselli *et al*. presented a cascade architecture containing a probabilistic mapping graph self-organizing map and a graph neural network to detect web spam [9]. The experimental results on WEBSPAM-UK2006 shown in Table III reached the state of the art at that time. Comparing the results with our experiment results (C4.5+FP+US), we can

find that the ensemble classifier integrated under-sampling and features partition methods outperforms the approaches presented by Scarselli *et al*. on F1-Measure and AUC metrics.

V. Conclusion

Web spam detection is an important topic in the field of information retrieval. Because web spam datasets are serious imbalance datasets, we propose an ensemble classifier based on under-sampling and features partition techniques. Experiments on WEBSPAM-UK2006 show that our method outperforms other approaches.

We plan to further our research by performing more experiments on other datasets and applications to test whether the ensemble approach based on features partition and under-sampling can be used generally. We also intend to integrate other algorithms such as KNN, Naive Bayesian etc. to verify whether the ensemble classifier proposed in this paper can improve their performance. Finally, the feature partition approach introduced in this paper is an empirical approach. We wish to propose a more theoretical features partition method.

Acknowledgment

The authors wish to acknowledge Carlos Castillo, who has supported the WEBSPAM-UK2006 Corpus web site and helped us to download the collection.

References

[1] Z. Gyöngyi and H. Garcia-Molina, "Web spam taxonomy," in *Proc. First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005)*, 2005.
[2] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a very large web search engine query log," in *Proc. SIGIR Forum*, 1999, vol. 33, no. 1, pp. 6-12.
[3] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proc. the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR' 05, Salvador, Brazil, 2005, pp. 154-161.
[4] N. Spirin and J. Han, "Survey on web spam detection: Principles and algorithms," *ACM SIGKDD Explorations Newsletter*, vol. 13, no. 2, pp. 50-64, 2012.
[5] L. Araujo and J. Martinez-Romo, "Web spam detection: New classification features based on qualified link analysis and language models," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 581-590, 2010.
[6] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates, "Web spam detection: Link-based and content-based techniques," in *Proc. the Final Workshop on The European Integrated Project Dynamically Evolving, Large Scale Information Systems (DELIS)*, 2008, pp. 99-113.
[7] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in *Proc. the 15th International Conference on World Wide Web. ACM*, 2006, pp. 83-92.
[8] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: Web spam detection using the web topology," in *Proc. the 30th Annual International ACM SIGIR, Conference on Research and Development in Information Retrieval ACM*, pp. 423-430, 2007.
[9] F. Scarselli, A. C. Tsoi, M. Hagenbuchner, and L. D. Noi, "Solving graph data issues using a layered architecture approach with applications to web spam detection," *Neural Networks*, vol. 48, pp. 78-90, 2013.
[10] J. S. Jegadeesh and P. L. Jacob, "Web spam detection using fuzzy clustering," *International Journal on Recent and Innovation Trends in Computing and Communication*, pp. 928- 938, 2013.
[11] W. Wang, X. D. Lee, A. L. Hu, and G. G. Geng, "Co-training based semi-supervised Web spam detection," in *Proc. 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2013, pp. 789-793.
[12] G. G. Geng, C. H. Wang, Q. D. Li, L. Xu, and X. B. Jin, "Boosting the performance of web spam detection with ensemble under-sampling classification," in *Proc. Fourth International Conference on Fuzzy Systems and Knowledge Discovery(FSKD)*, 2007, vol. 4, pp. 583-587.
[13] Y. H. Peng, "A novel ensemble machine learning for robust microarray data classification," *Computers in Biology and Medicine*, vol. 36, no. 6, pp. 553-573, 2006.
[14] X. Liu, J. Wu, and Z. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 539-550, 2009.
[15] C. Castillo, D. Donato, L. Becchetti, P. Boldi, M. Santini, and S., Vigna. "A reference collection for web spam," in *Proc. SIGIR Forum*, vol. 40, no. 2, pp. 11-24, 2006.
[16] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*, Cambridge: Cambridge University Press, 2008.

Xiaoyong Lu received his BA degree and Ph.D. degree in management science and engineering from Tianjin University. Currently, he is a professor in Nanchang University. His major interests include data mining and knowledge discovery, information management and information system etc. He published several highly cited papers in journals and conferences which are indexed by EI.

Musheng Chen received his master's degree in software engineering in Information Engineering School of Nanchang University. Currently, he is a doctoral candidate of information management and information system in Information Engineering School of Nanchang University and College of Informatics in Yuan Ze University. His major interests include data mining and knowledge discovery, information management and information system etc.

Jhenglong Wu received his Ph.D. degree in the Department of Information Management, Yuan Ze University. Currently, he is a postdoctoral fellow in Academia Sinica in Taiwan. His major interests include data mining, text mining, sentimental analysis, technology management and applications of soft computing. He has published papers in Journals of Applied Soft Computing, Soft Computing etc.

Peichann Chang received his BA degree in industrial engineering from National Tsing Hua University in 1979, MS and PhD degrees from the Department of Industrial Engineering of Lehigh University in 1985 and 1989, respectively. Currently, he is a chair professor in Yuan Ze University, Taiwan. He has been rewarded as the outstanding researcher from National Research Council from 2009 to 2011. He published several highly cited papers in journals such as Applied Soft Computing, Journal of Intelligent Manufacturing and Neurocomputing. He is in the editorial board of Applied Soft Computing and also serves as a referee for more than 30 international journals. His research interests include evolutionary computations, financial time series data forecasting, medical data classification and diagnosis, fuzzy rules based systems, production scheduling, and applications of soft computing. His publications have the total citations of 4155 and an h-index of 36 from Google Scholar database.